Task Set 1 — Interpreting Algorithmic Functional Dependencies

1. Objective

The goal of Task Set 1 is to **analyze and interpret the functional dependencies (FDs)** produced by an automatic FD discovery algorithm (TANE), **without reasoning about their semantic validity**. The objective is to understand:

- how many FDs are produced per dataset,
- the structure of these FDs (size of determinants),
- which attributes frequently appear on the left-hand side (LHS) and right-hand side (RHS),
- and which patterns appear suspicious or trivial from a purely algorithmic perspective.

No FD discovery was performed in this task; all FDs were provided.

---

2. Dataset Overview

Dataset	**Rows**	**Cols**	**#FDs**
iris	150	5	4
balance-scale	625	5	1
chess	28,056	7	1
abalone	4,177	9	137
nursery	12,960	9	1
breast-cancer-wisconsin	699	11	46
bridges	108	13	142
echocardiogram	132	13	538
hepatitis	155	20	8,250
adult	48,842	14	78 (table)
horse	300	27	128,726 (table)

---

3. Per-Dataset Analysis

3.1 Iris

- **FD count:** 4
- **Average LHS size:** 3
- **FD structure:** All FDs are of the form
  {three measurements} → species
- **Observations:**
  - No single attribute or attribute pair determines the class.
  - All four measurement attributes appear symmetrically on the LHS.
- **Interpretation:**
  The algorithm identifies strong geometric separability between species, producing multiple equivalent determinants.

3.2 Balance-Scale

- **FD count:** 1
- **Average LHS size:** 4
- **FD:**
  {left weight, left distance, right weight, right distance} → class
- **Observations:**
  - All configuration attributes are required.
  - No smaller determinant exists.
- **Interpretation:**
  This reflects a deterministic physical rule encoded in the dataset.

3.3 Chess

- **FD count:** 1
- **Average LHS size:** 6
- **FD:**
  {all position attributes} → outcome

- **Observations:**
  - The full board configuration is required to determine the result.
- **Interpretation:**
  The dataset encodes a deterministic game state; partial descriptions are insufficient.

3.4 Nursery

- **FD count:** 1
- **Average LHS size:** 8
- **FD:**
  {all application attributes} → class
- **Observations:**
  - The dataset is rule-based and synthetic.
  - No subset of attributes determines the decision.
- **Interpretation:**
  The FD reflects a designed decision function rather than empirical correlations.

3.5 Abalone

- **FD count:** 137
- **Average LHS size:** ≈ 4.2
- **Dominant LHS attributes:** weight-related measurements
- **Observations:**
  - Many large determinants (size 4–6).
  - The same LHS often determines multiple unrelated RHS attributes.
- **Interpretation:**
  Continuous numeric attributes combine to form quasi-identifiers, leading to many accidental FDs.

3.6 Breast-Cancer-Wisconsin

- **FD count:** 46

- **Average LHS size:** ≈ 2.8
- **Key patterns:**
  - Identifier-based FDs: SampleID → X
  - Measurement combinations determining diagnosis
- **Observations:**
  - Presence of both trivial and non-trivial dependencies.
- **Interpretation:**
  The dataset exhibits a mix of identifier-driven and correlation-driven FDs.

3.7 Bridges

- **FD count:** 142
- **Average LHS size:** ≈ 3.6
- **Key patterns:**
  - Bridge ID determining most attributes
  - Strong correlations between material, type, and span length
- **Observations:**
  - Very high FD count relative to dataset size.
  - Many large determinants.
- **Interpretation:**
  The small dataset size and categorical redundancy lead to FD explosion.

3.8 Echocardiogram

- **FD count:** 538
- **Average LHS size:** ≈ 4.9
- **Observations:**
  - Numerous large determinants.
  - Many numeric attributes act as quasi-identifiers.

- **Interpretation:**
  High-precision medical measurements in a small dataset produce extensive overfitting by FD discovery.

3.9 Hepatitis

- **FD count:** 8,250
- **Average LHS size:** ≈ 5.6
- **Key factors:**
  - Many binary attributes
  - Extensive missing values (?)
- **Observations:**
  - Extreme FD explosion.
  - Missing values collapse variability and create artificial determinism.
- **Interpretation:**
  This dataset highlights the severe limitations of exact FD discovery on sparse medical data.

3.10 Adult (Special Case)

- **FD count:** 78 (from assignment table)
- **Provided FD file:** empty
- **Interpretation:**
  - The absence of FDs in the provided file is likely due to preprocessing or missing-value handling.
  - The discrepancy illustrates the sensitivity of FD discovery to data representation.
- **Conclusion:**
  The Adult dataset exemplifies realistic, noisy data where exact global constraints are difficult to extract.

3.11 Horse (Special Case)

- **FD count:** 128,726 (from assignment table)

- **Provided FD file:** empty
- **Key characteristics:**
  - 27 attributes
  - Many missing values
  - Mixed numeric and categorical data
- **Interpretation:**
  The dataset theoretically produces an extreme number of FDs, but practical extraction is hindered by missing values and tooling limitations.

---

4. Cross-Dataset Comparison

Dataset Type	**Typical Behavior**
Synthetic / rule-based (Balance-Scale, Nursery, Chess)	Single large FD
Small numeric (Abalone, Echocardiogram)	Many large accidental FDs
Medical with missing values (Hepatitis, Horse)	Extreme FD explosion
Realistic noisy data (Adult)	Few or no exact FDs